A Review of Stochastic Gradient Descent

Ravi Raghavan & Lakshitha Ramanayake

March 2024

1 Stochastic Gradient Descent Basic Concepts

1.1 Theory: Smooth Functions and Convexity

When discussing convergence, we will first discuss the underlying theory of Smooth Functions and Convexity.

1.1.1 Differentiability

Definition 1 (Jacobian). Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^p$ be differentiable, and $x \in \mathbb{R}^d$. Then, we note $\mathcal{DF}(x)$ the **Jacobian** of $\mathcal{F}$ at x, which is the matrix defined by its first partial derivatives:

$$[\mathcal{DF}(x)]_{ij} = \frac{\partial f_i}{\partial x_j}(x), \quad for \;\; i = 1, \ldots, p, j = 1, \ldots, d \tag{1}$$

Remark 2 (Gradient). If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then $\mathcal{D}f(x) \in \mathbb{R}^{1xd}$ is a row vector, whose transpose is called the **gradient** of f at x: $\nabla f(x) = \mathcal{D}f(x)^T \in \mathbb{R}^{dx1}$

Definition 3 (Hessian). Let $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, and $x \in \mathbb{R}^d$. Then we note $\nabla^2 f(x)$ the **Hessian** of f at x, which is the matrix defined by its second-order partial derivatives:

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \quad for \;\; i, j = 1, \ldots, d \tag{2}$$

Consequently $\nabla^2 f(x)$ is a $d \;\; x \;\; d$ matrix.

Definition 4 (Lipschitz). Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^p$, and $L > 0$. We say that F is L-**Lipschitz** if

$$\forall x, y \in \mathbb{R}^d, \;\; ||\mathcal{F}(y) - \mathcal{F}(x)|| \leq L||y - x|| \tag{3}$$

1.1.2 Convexity

Definition 5 (Jensen's Inequality). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if **dom** f is a convex set and if for all $x, y \in domf$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \tag{4}$$

Definition 6 (First Order Condition of Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if and only if **dom** f is a convex set and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{5}$$

holds for all $x, y \in domf$

Definition 7 (Second Order Condition of Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if and only if **dom** f is a convex set and its Hessian is positive semi-definite: for all $x \in domf$,

$$\nabla^2 f(x) \succeq 0 \tag{6}$$

1.1.3 Strong Convexity

Definition 5 (Jensen's Inequality for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if **dom** f is a convex set and if for all $x, y \in domf$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{p}{2}\theta(1-\theta)||x-y||_2^2 \tag{7}$$

Definition 6 (First Order Condition for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if and only if **dom** f is a convex set and

$$f(y) \geq f(x) + \nabla(f(x))^T(y-x) + \frac{p}{2}||y-x||_2^2 \tag{8}$$

holds for all $x, y \in domf$

Definition 7 (Second Order Condition for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if and only if **dom** f is a convex set and

$$\nabla^2 f(x) \succeq pI \tag{9}$$

1.1.4 Smoothness

Definition 8 (L-Smooth Functions). A function $f : \mathbb{R}^n \to \mathbb{R}$ and $L > 0$ is L-Smooth if it is differentiable and if $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is $L-$Lipschitz:

$$\forall x, y \in \mathbb{R}^n, \;\; ||\nabla f(y) - \nabla f(x)|| \leq L||y-x|| \tag{10}$$

As a subsequent node, $L-$Smooth functions have a quadratic upper bound:

$$\forall x, y \in \mathbb{R}^n, f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}||y-x||^2 \tag{11}$$

Lemma: If f is $L-$smooth and $\gamma > 0$ then,

$$\forall x, y \in \mathbb{R}^n, \;\; f(x - \gamma \nabla f(x)) - f(x) \leq -\gamma(1 - \frac{\gamma L}{2})||\nabla f(x)||^2 \tag{12}$$

Proof:
Let's begin with the definition of $L-$Smooth. Earlier, we showed that when a function is $L-$Smooth, the following holds true:

$||\nabla f(x) - \nabla f(y)|| \leq L||x-y||$

Let's now begin with the proof:
Let $g(t) = f(x + t(y-x))$

Based on simple calculus rules
$f(y) - f(x) = \int_0^1 g'(t)\, dt$

$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))^T (y-x)\, dt$

$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y-x)), y-x \rangle\, dt$

$f(y) - f(x) = \langle \nabla f(x), y-x \rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y-x \rangle\, dt$

Apply the Cauchy Schwarz Inequality on the inside of the integral

$f(y) - f(x) \leq \langle \nabla f(x), y-x \rangle + \int_0^1 ||\nabla f(x + t(y-x)) - \nabla f(x)||||y-x||\, dt$

Now let's apply the definition of $L-$Smoothness inside the integral,

$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \int_0^1 tL||y - x||^2 \, dt$

$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}||y - x||^2$

Now, let's insert $y = x - \gamma \nabla f(x)$ into the above equation

$f(x - \gamma \nabla f(x)) - f(x) \leq \langle \nabla f(x), x - \gamma \nabla f(x) - x \rangle + \frac{L}{2}||x - \gamma \nabla f(x) - x||^2$

$f(x - \gamma \nabla f(x)) - f(x) \leq \langle \nabla f(x), -\gamma \nabla f(x) \rangle + \frac{L}{2}|| - \gamma \nabla f(x)||^2$

$f(x - \gamma \nabla f(x)) - f(x) \leq -\langle \nabla f(x), \gamma \nabla f(x) \rangle + \frac{L\gamma^2}{2}||\nabla f(x)||^2$

$f(x - \gamma \nabla f(x)) - f(x) \leq -\gamma||\nabla f(x)||^2 + \frac{L\gamma^2}{2}||\nabla f(x)||^2$

$f(x - \gamma \nabla f(x)) - f(x) \leq (-\gamma + \frac{L\gamma^2}{2})||\nabla f(x)||^2$

If we assume that $\inf f > -\infty$ and if we set $\gamma = \frac{1}{L}$, we can see that:

$\inf f - f(x) \leq f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq (-\frac{1}{2L})||\nabla f(x)||^2$

$(\frac{1}{2L})||\nabla f(x)||^2 \leq f(x) - \inf f$

1.1.5 Smoothness and Convexity

Lemma: If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and $L-$smooth, then $\forall x, y \in \mathbb{R}^d$, we have that:

$$\frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \tag{13}$$

Proof: $f(x) - f(y) = f(x) - f(z) + f(z) - f(y)$

We know that, due to the first order condition of convexity, $f(z) \geq f(x) + \nabla f(x)^T(z - x)$
$f(x) - f(z) \leq -\nabla f(x)^T(z - x) = \nabla f(x)^T(x - z)$
We also know that, due to the quadratic upper bound property of L-Smooth functions,
$f(z) \leq f(y) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}||z - y||^2$
$f(z) - f(y) \leq \langle \nabla f(y), z - y \rangle + \frac{L}{2}||z - y||^2$

Hence, $f(x) - f(y) = f(x) - f(z) + f(z) - f(y) \leq \nabla f(x)^T(x - z) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}||z - y||^2$
Let's aim to minimize the right hand side.
Let's take the derivative of the right hand side with respect to z. This gradient is:
$-\nabla f(x) + \nabla f(y) + \frac{L}{2}(2z - 2y)$
$-\nabla f(x) + \nabla f(y) + L(z - y)$
Setting this to zero gives us: $z - y = \frac{1}{L}(\nabla f(x) - \nabla f(y))$
$z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x))$
Now that we have found the quantity to minimize the right hand side, we can substitute it in.
$\nabla f(x)^T(x - (y - \frac{1}{L}(\nabla f(y) - \nabla f(x)))) + \langle \nabla f(y), y - \frac{1}{L}(\nabla f(y) - \nabla f(x)) - y \rangle + \frac{L}{2}||y - \frac{1}{L}(\nabla f(y) - \nabla f(x)) - y||^2$

$\nabla f(x)^T(x - (y - \frac{1}{L}(\nabla f(y) - \nabla f(x)))) - \langle \nabla f(y), \frac{1}{L}(\nabla f(y) - \nabla f(x)) \rangle + \frac{L}{2}||\frac{1}{L}(\nabla f(y) - \nabla f(x))||^2$

$\nabla f(x)^T(x - y) + \frac{1}{L}\nabla f(x)^T(\nabla f(y) - \nabla f(x)) - \langle \nabla f(y), \frac{1}{L}(\nabla f(y) - \nabla f(x)) \rangle + \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$

$\nabla f(x)^T(x - y) + \frac{1}{L}\langle \nabla f(x) - \nabla f(y), \nabla f(y) - \nabla f(x) \rangle + \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$

$\nabla f(x)^T(x - y) - \frac{1}{L}\langle \nabla f(y) - \nabla f(x), \nabla f(y) - \nabla f(x) \rangle + \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$

$\nabla f(x)^T(x - y) - \frac{1}{L}||\nabla f(y) - \nabla f(x)||^2 + \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$

$\nabla f(x)^T(x-y) - \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$

We now see that $f(x) - f(y) \leq \nabla f(x)^T(x-y) - \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$
Rearranging terms will make us see that $\frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$

1.2 Gradient Descent

Gradient Descent Algorithm. Let $x^{(0)} \in \mathbb{R}^d$, and let $\gamma > 0$ be a step size. The **Gradient Descent (GD)** algorithm defines a sequence $(x^{(t)})_{t \in \mathbb{N}}$ satisfying

$$x^{(t+1)} = x^{(t)} - \gamma \nabla f(x^{(t)}) \tag{14}$$

1.3 Function Definitions

Sum of Functions. Let's say that we have a function $f : \mathbb{R}^d \to \mathbb{R}$ which can be expressed as:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{15}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$. We can say that f is a Sum of Functions
Sum of Convex Functions. Let's say that we have a function $f : \mathbb{R}^d \to \mathbb{R}$ which can be expressed as:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{16}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ and f_i is convex. We can say that f is a Sum of Convex Functions
Sum of L-Smooth Functions. Let's say that we have a function $f : \mathbb{R}^d \to \mathbb{R}$ which can be expressed as:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{17}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ and f_i is L_i smooth. We can say that f is a Sum of L-Smooth Functions. Let $L_{max} = \max_{1,\dots,n} L_i$

1.4 Stochastic Gradient Descent

Stochastic Gradient Descent Algorithm. Let's say that we are minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ which is a sum of functions. It is assumed that $\arg\min f \neq \emptyset$ and that f_i is unbounded below.
Let $x^{(0)} \in \mathbb{R}^d$, and let $\gamma_t > 0$ be a sequence of step sizes. The **Stochastic Gradient Descent (GD)** algorithm defines a sequence $(x^{(t)})_{t \in \mathbb{N}}$ satisfying

$$i_t \in \{1, \dots, n\} \tag{18}$$

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{i_t}(x^{(t)}) \tag{19}$$

Note i_t is sampled with probability $\frac{1}{n}$
It is evident to see that the gradients used during Stochastic Gradient Descent is an unbiased estimator of $\nabla f(x)$. $E[\nabla f_i(x)] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) = \nabla f(x)$

1.5 Minibatch Stochastic Gradient Descent

Minibatch Stochastic Gradient Descent Algorithm. Let's say that we are minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ which is a sum of functions. It is assumed that $\arg\min f \neq \emptyset$ and that f_i is unbounded below.
Let $x^{(0)} \in \mathbb{R}^d$, let $b \in [1, n]$ be the batch size, and let $\gamma_t > 0$ be a sequence of step sizes. The **Minibatch Stochastic Gradient Descent (Minibatch SGD)** algorithm defines a sequence $(x^{(t)})_{t \in \mathbb{N}}$ satisfying

$$B_t \subset \{1, \dots, n\} \tag{20}$$

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{B_t}(x^{(t)}) \tag{21}$$

Note B_t is sampled uniformly among all sets of size b. This means that given a batch of size b, it has a probability of $\frac{1}{\binom{n}{b}}$ of being selected

$$\nabla f_{B_t}(x^{(t)}) = \frac{1}{|B|}\sum_{i\in B}\nabla f_i(x^{(t)}) \tag{22}$$

It is easy to see that:

$$\mathbb{E}[\nabla f_{B_t}(x^{(t)})] = \frac{1}{\binom{n}{b}}\sum_{B\subset\{1,\dots,n\},|B|=b}\nabla f_B(x^{(t)}) \tag{23}$$

We can easily prove that

$$\mathbb{E}[\nabla f_{B_t}(x^{(t)})] = \nabla f(x^{(t)}) \tag{24}$$

The key observation we must make is that $\nabla f_i(x^{(t)})$ will be used when computing mini batch gradients(i.e. $\nabla f_{B_t}(x^{(t)})$) in exactly $\binom{n-1}{b-1}$ mini batches.

Combinatoric Property: $\binom{n}{b} = \binom{n-1}{b-1}\cdot\frac{n}{b}$

$$\mathbb{E}[\nabla f_{B_t}(x^{(t)})] = \frac{1}{\binom{n}{b}}\binom{n-1}{b-1}\sum_{i=1}^{n}\frac{1}{b}\nabla f_i(x^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^{(t)}) = \nabla f(x^{(t)}) \tag{25}$$

1.6 Expected Smoothness and Variance

1.6.1 Expected Smoothness

The goal of this section is to analyze the "expected properties" of f_i.

Lemma. Let's say that we have a function $f:\mathbb{R}^d\to\mathbb{R}$ which is a sum of Convex functions and a sum of L-Smooth functions, we can state the following:

$$\forall x,y\in\mathbb{R}^d,\ \ \frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(y)-\nabla f_i(x)||^2] \le f(y)-f(x)-\langle\nabla f(x),y-x\rangle \tag{26}$$

Proof: We can use the Lemma that we derived in Section 1.1.5 as well as the fact that $L_i\le L_{max}$ to see the following:

$\frac{1}{2L_{max}}||\nabla f_i(y)-\nabla f_i(x)||^2 \le \frac{1}{2L_i}||\nabla f_i(y)-\nabla f_i(x)||^2 \le f_i(y)-f_i(x)-\langle\nabla f_i(x),y-x\rangle$

$\frac{1}{2L_{max}}||\nabla f_i(y)-\nabla f_i(x)||^2 \le f_i(y)-f_i(x)-\langle\nabla f_i(x),y-x\rangle$

$\frac{1}{n}(\frac{1}{2L_{max}}||\nabla f_i(y)-\nabla f_i(x)||^2) \le \frac{1}{n}(f_i(y)-f_i(x)-\langle\nabla f_i(x),y-x\rangle)$

We are now prepared to take expectation. The above inequality can be duplicated for each value of i and we can sum these inequalities across all values of i. Hence, we see the following:

$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(y)-\nabla f_i(x)||^2] \le f(y)-f(x)-\langle\nabla f(x),y-x\rangle$
If we set $x=x^*$ and $y=x$, we can see that
$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x)-\nabla f_i(x^*)||^2] \le f(x)-f(x^*)-\langle\nabla f(x^*),x-x^*\rangle$

$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x)-\nabla f_i(x^*)||^2] \le f(x)-inff$

1.6.2 Variance

It is rather intuitive that, given a particular value of x, if the values of f_i at this value of x exhibit a lower variance, this would indicate that our stochastic algorithm would converge faster. This section is going to dive deep into this variance concept and provide a few measurable quantities that will quantify this concept.

Definition. Let's say we have a function that is a Sum of Functions. **Interpolation** holds if there exists a common $x^*\in\mathbb{R}^d$ such that $f_i(x^*)=inff_i, \forall i=1,\dots n$. We will state that interpolation occurs at x^*

Lemma. Let's say we have a function that is a Sum of Functions. If interpolation holds at $x^*\in\mathbb{R}^d$, then $x^*\in\arg\min f$

Proof. Since we know that interpolation holds at x^*, we know that $f_i(x^*)=inff_i, \forall i=1,\dots n$

$f(x^*)=\frac{1}{n}\sum_{i=1}^n f_i(x^*)=\frac{1}{n}\sum_{i=1}^n inff_i \le \frac{1}{n}\sum_{i=1}^n f_i(x)=f(x)$

Definition. Again, we are looking at functions that are Sum of Functions. Let us define a new quantity called

function noise.

$$\Delta_f^* = inff - \frac{1}{n}\sum_{i=1}^{n} inff_i \tag{27}$$

Lemma. Given the previous definition, we can state the following

$$\Delta_f^* \geq 0 \tag{28}$$

Interpolation Holds if and only if $\Delta_f^* = 0$

Proof. $\Delta_f^* = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} inff_i \geq f(x^*) - \frac{1}{n}\sum_{i=1}^{n} f_i(x^*) = f(x^*) - f(x^*) = 0$

Let's now prove that Interpolation Holds if and only if $\Delta_f^* = 0$.

First we will prove the first direction. So let it be the case that interpolation holds.

When interpolation holds, $inff_i = f_i(x^*)$. Hence, we can say that $\Delta_f^* = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} inff_i = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} f_i(x^*) = f(x^*) - f(x^*) = 0$

Now let's prove the second direction. Let it be the case that $\Delta_f^* = 0$.

We know that $\Delta_f^* = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} inff_i \geq f(x^*) - \frac{1}{n}\sum_{i=1}^{n} f_i(x^*) = f(x^*) - f(x^*) = 0$

This means that $\sum_{i=1}^{n} inff_i = \sum_{i=1}^{n} f_i(x^*)$

$\sum_{i=1}^{n} f_i(x^*) - inff_i = 0$

We know that $f_i(x^*) \geq inff_i$. Hence, for the above equality to hold true, we must observe that $f_i(x^*) = inff_i$ which means that Interpolation must hold

Definition Let us now work with a function that is a Sum of L-Smooth Functions. We will define a term, called **gradient noise** that is set up as follows:

$$\sigma_f^* = \inf_{x^* \in \arg\min f} V[\nabla f_i(x^*)] \tag{29}$$

where $V[X] = E[||X - E[X]||^2]$

Definition: Let us now work with a function that is a Sum of L-Smooth Functions. We will define a term, called **minibatch gradient noise** that is set up as follows:

$$\sigma_b{}^* = \inf_{x^* \in \arg\min f} V[\nabla f_B(x^*)] \tag{30}$$

Definition: Let us now work with a function that is a Sum of L-Smooth Functions. Let $b \in [1, n]$. We can say that f is L_b smooth in expectation if

$$\forall x, y \in \mathbb{R}^d, \frac{1}{2L_b}\mathbb{E}[||\nabla f_B(y) - \nabla f_B(x)||^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \tag{31}$$

If $y = x$ and $x = x^*$, we can see that a function being L_b smooth indicates that:

$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle$

$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq f(x) - \inf f$

Lemma. Let us say we have a function that is a sum of L-Smooth Functions.

If we also have that this function is a sum of Convex functions, we can claim that

$$\sigma_f^* = V[\nabla f_i(x^*)], \forall x^* \in \arg\min f \tag{32}$$

Proof. Let us denote $x_1, x^* \in \arg\min f$. If we can show that $V[\nabla f_i(x_1)] = V[\nabla f_i(x^*)]$, then we will have completed the proof

$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x_1) - \nabla f_i(x^*)||^2] \leq f(x_1) - f(x^*) - \langle \nabla f(x^*), x_1 - x^* \rangle$

We know that $f(x^*) = inff$ and that $\nabla f(x^*) = 0$

$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x_1) - \nabla f_i(x^*)||^2] \leq f(x_1) - inff = 0$

Since $||\nabla f_i(x_1) - \nabla f_i(x^*)||^2$ is a positive quantity, this must mean that $\mathbb{E}[||\nabla f_i(x_1) - \nabla f_i(x^*)||^2] = 0$ and that $||\nabla f_i(x_1) - \nabla f_i(x^*)|| = 0$ for all i.

Hence, we have shown that $f_i(x_1) = f_i(x^*)$ and that, subsequently, $V[\nabla f_i(x_1)] = V[\nabla f_i(x^*)]$

Lemma. Let us say we have a function that is a sum of L-Smooth Functions.

If we also have that this function is a sum of Convex functions, we can claim that

$$\sigma_b^* = V[\nabla f_B(x^*)], \forall x^* \in \arg\min f \tag{33}$$

Proof. Let us denote $x_1, x^* \in \arg\min f$. If we can show that $V[\nabla f_B(x_1)] = V[\nabla f_B(x^*)]$, then we will have completed the proof

$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x^*)||^2] \leq f(x_1) - f(x^*) - \langle \nabla f(x^*), x_1 - x^* \rangle$

We know that $f(x^*) = inf f$ and that $\nabla f(x^*) = 0$

$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x^*)||^2] \leq f(x_1) - inf f = 0$

Since $||\nabla f_B(x_1) - \nabla f_B(x^*)||^2$ is a positive quantity, this must mean that $\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x^*)||^2] = 0$ and that $||\nabla f_B(x_1) - \nabla f_B(x^*)|| = 0$ for all i.

Hence, we have shown that $f_B(x_1) = f_B(x^*)$ and that, subsequently, $V[\nabla f_B(x_1)] = V[\nabla f_B(x^*)]$

From the aforementioned analysis it is clear that both the function noise and gradient noise measure how close/far away from interpolation we are.

Lemma. Let us say that we have a function that is a Sum of L-Smooth Functions.

1. $\sigma_f^* \leq 2L_{max}\Delta_f^*$
2. If each f_i is p strongly convex, then $2p\Delta_f^* \leq \sigma_f^*$

Proof: Let us say that we have a point $x^* \in \arg\min f$. Earlier, we show that if a function f is $L-$Smooth, then $||\nabla f(x)||^2 \leq 2L * (f(x) - \inf f)$.

In this case, we know that each f_i is $L-$Smooth. We can proceed as follows:
$||\nabla f_i(x^*)||^2 \leq 2L_i * (f_i(x^*) - \inf f_i) \leq 2L_{max} * (f_i(x^*) - \inf f_i)$.

Let's take the expectation over both sides of the inequality:
$\mathbb{E}[||\nabla f_i(x^*)||^2] \leq \mathbb{E}[2L_{max} * (f_i(x^*) - \inf f_i)]$.

$\mathbb{E}[||\nabla f_i(x^*)||^2] \leq 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)]$.

Let's take a look at $\mathbb{E}[||\nabla f_i(x^*)||^2]$. We will use the fact that $\nabla f(x^*) = 0$

$\mathbb{E}[||\nabla f_i(x^*)||^2] = \mathbb{E}[||\nabla f_i(x^*) - \nabla f(x^*)||^2] = \mathbb{E}[||\nabla f_i(x^*) - \mathbb{E}[\nabla f_i(x^*)]||^2] = \mathbb{V}[\nabla f_i(x^*)] \geq \sigma_f^*$

Now, let's analyze $2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)]$.

$2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] = 2L_{max}(\mathbb{E}[(f_i(x^*))] - \mathbb{E}[(\inf f_i)]) = 2L_{max}(f(x^*) - \frac{1}{n}\sum_{i=1}^n inf f_i)$

This can be simplified further:
$2L_{max}(\inf f - \frac{1}{n}\sum_{i=1}^n \inf f_i) = 2L_{max}\Delta_f^*$

We have shown that $\sigma_f^* \leq 2L_{max}\Delta_f^*$

Now, our job is to show that If each f_i is p strongly convex, then $2p\Delta_f^* \leq \sigma_f^*$

The definition of strong convexity tells us that, when a function f is p strongly convex and $x^* \in \arg\min f$:
$f(x) - f(x^*) \leq \frac{1}{2p}||\nabla f(x)||^2$

When a function f_i is p strongly convex, we have the inequality for each f_i:
$f_i(x) - \inf f_i \leq \frac{1}{2p}||\nabla f_i(x)||^2$

$f_i(x^*) - \inf f_i \leq \frac{1}{2p}||\nabla f_i(x^*)||^2$

Take Expectation over this inequality:

$\mathbb{E}[f_i(x^*) - \inf f_i] \leq \frac{1}{2p}\mathbb{E}[||\nabla f_i(x^*)||^2]$

$\mathbb{E}[f_i(x^*)] - \mathbb{E}[\inf f_i] \leq \frac{1}{2p}\mathbb{E}[||\nabla f_i(x^*)||^2]$

$\inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i \leq \frac{1}{2p}\mathbb{V}[\nabla f_i(x^*)]$

$\Delta_f^* \leq \frac{1}{2p}\mathbb{V}[\nabla f_i(x^*)]$

Since we have convexity, we know that $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$

$\Delta_f^* \leq \frac{1}{2p}\sigma_f^*$

$2p\Delta_f^* \leq \sigma_f^*$

Lemma. Let us say that we have a function that is a Sum of L-Smooth Functions.

$$\forall x \in \mathbb{R}^d, \;\; \mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\Delta_f^* \tag{34}$$

Proof: Earlier, we showed that, for L-Smooth Functions, $(\frac{1}{2L})||\nabla f(x)||^2 \leq f(x) - \inf f$

Hence, we can state that: $(\frac{1}{2L_i})||\nabla f_i(x)||^2 \leq f_i(x) - \inf f_i$

$||\nabla f_i(x)||^2 \leq 2L_i(f_i(x) - \inf f_i) \leq 2L_{max}(f_i(x) - \inf f_i) = 2L_{max}(f_i(x) - f_i(x^*)) + 2L_{max}(f_i(x^*) - \inf f_i)$

$||\nabla f_i(x)||^2 \leq 2L_{max}(f_i(x) - f_i(x^*)) + 2L_{max}(f_i(x^*) - \inf f_i)$

Let's take expectation
$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}\mathbb{E}[(f_i(x) - f_i(x^*))] + 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)]$

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)]$

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}(\inf f - E[\inf f_i])$

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}(\inf f - \frac{1}{n}\sum_{i=1}^{n} inf f_i)$

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\Delta_f^*$

Lemma. Let us say that we have a function that is a Sum of L-Smooth and Convex Functions.

$$\forall x \in \mathbb{R}^d, \;\; \mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^* \tag{35}$$

Proof:
For the rest of this proof, let's pick an $x^* \in \arg\min f$
In our journey to prove this lemma, the first thing we can do is express $||\nabla f_i(x)||^2$ as follows:

$||\nabla f_i(x)||^2 = ||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2$

Based on the triangular inequality for norms, we know that $||x + y|| \leq ||x|| + ||y||$

$||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \leq ||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)|| + ||\nabla f_i(x^*)||^2$

We know that $||\nabla f_i(x) - \nabla f_i(x^*) - f_i(x^*)||^2 \geq 0$

$||\nabla f_i(x) - \nabla f_i(x^*)||^2 - 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)|| + ||\nabla f_i(x^*)||^2 \geq 0$

$||\nabla f_i(x) - \nabla f_i(x^*)||^2 + ||\nabla f_i(x^*)||^2 \geq 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)||$

We can express $||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \leq ||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)|| + ||\nabla f_i(x^*)||^2$ as follows:

$||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \leq 2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2$

$||\nabla f_i(x)||^2 \leq 2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2$

The next thing we can do is take the expectation over this inequality

$\mathbb{E}[||\nabla f_i(x)||^2] \leq \mathbb{E}[2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2]$

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2 + ||\nabla f_i(x^*)||^2]$

Earlier, we showed that $\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq f(x) - inf f$

This basically means that $\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq 2L_{max}(f(x) - \inf f)$

$2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq 4L_{max}(f(x) - \inf f)$

Since we have a function that is a Sum of L-Smooth and Convex Functions, we know that $\sigma_f^* = V[\nabla f_i(x^*)], \forall x^* \in \arg\min f$.

$\sigma_f^* = E[||\nabla f_i(x^*) - E[\nabla f_i(x^*)]||^2], \forall x^* \in \arg\min f.$

We know that $E[\nabla f_i(x^*)] = \nabla f(x^*) = 0$ since f is a Sum of Convex Functions.

In this case, $\sigma_f^* = E[||\nabla f_i(x^*)||^2]$

$2\sigma_f^* = 2E[||\nabla f_i(x^*)||^2]$

According to Linearity of Expectation, we know that $2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2 + ||\nabla f_i(x^*)||^2] = 2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] + 2\mathbb{E}[||\nabla f_i(x^*)||^2]$

Using this fact, we can further simplify this inequality: $\mathbb{E}[||\nabla f_i(x)||^2] \leq \mathbb{E}[2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2]$

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] + 2\mathbb{E}[||\nabla f_i(x^*)||^2]$

Substituting the values that we derived earlier(i.e. $2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq 4L_{max}(f(x) - \inf f)$ and $2\sigma_f^* = 2E[||\nabla f_i(x^*)||^2]$)

$\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$

Lemma Let us have a function that is a sum of L-Smooth Functions and a sum of Convex functions.

$$\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^* \tag{36}$$

Proof:

For the rest of this proof, let's pick an $x^* \in \arg\min f$

In our journey to prove this lemma, the first thing we can do is express $||\nabla f_B(x)||^2$ as follows:

$||\nabla f_B(x)||^2 = ||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2$

Based on the triangular inequality for norms, we know that $||x + y|| \leq ||x|| + ||y||$

$||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2 = ||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x) - \nabla f_B(x^*)||||\nabla f_B(x^*)|| + ||\nabla f_B(x^*)||^2$

We know that $||\nabla f_B(x) - \nabla f_B(x^*) - f_B(x^*)||^2 \geq 0$

$||\nabla f_B(x) - \nabla f_B(x^*)||^2 - 2||\nabla f_B(x) - \nabla f_B(x^*)||||\nabla f_B(x^*)|| + ||\nabla f_B(x^*)||^2 \geq 0$

$||\nabla f_B(x) - \nabla f_B(x^*)||^2 + ||\nabla f_B(x^*)||^2 \geq 2||\nabla f_B(x) - \nabla f_B(x^*)||||\nabla f_B(x^*)||$

We can express $||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2 = ||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x) - \nabla f_B(x^*)||||\nabla f_B(x^*)|| + ||\nabla f_B(x^*)||^2$ as follows:

$||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2 \leq 2||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x^*)||^2$

$||\nabla f_B(x)||^2 \leq 2||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x^*)||^2$

The next thing we can do is take the expectation over this inequality

$\mathbb{E}[||\nabla f_B(x)||^2] \leq \mathbb{E}[2||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x^*)||^2]$

$\mathbb{E}[||\nabla f_B(x)||^2] \leq 2\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2 + ||\nabla f_B(x^*)||^2]$

Earlier, we showed that $\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq f(x) - inf f$

This basically means that $\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq 2L_b(f(x) - \inf f)$

$2\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq 4L_b(f(x) - \inf f)$

Since we have a function that is a Sum of L-Smooth and Convex Functions, we know that $\sigma_b^* = V[\nabla f_B(x^*)], \forall x^* \in \arg\min f$.

$\sigma_b^* = E[||\nabla f_B(x^*) - E[\nabla f_B(x^*)]||^2], \forall x^* \in \arg\min f$.

We know that $E[\nabla f_B(x^*)] = \nabla f(x^*) = 0$ since f is a Sum of Convex Functions.

In this case, $\sigma_b^* = E[||\nabla f_B(x^*)||^2]$

$2\sigma_b^* = 2E[||\nabla f_B(x^*)||^2]$

According to Linearity of Expectation, we know that $2\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2 + ||\nabla f_B(x^*)||^2] = 2\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] + 2\mathbb{E}[||\nabla f_B(x^*)||^2]$

Using this fact, we can further simplify this inequality: $\mathbb{E}[||\nabla f_B(x)||^2] \leq \mathbb{E}[2||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x^*)||^2]$

$\mathbb{E}[||\nabla f_B(x)||^2] \leq 2\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] + 2\mathbb{E}[||\nabla f_B(x^*)||^2]$

Substituting the values that we derived earlier(i.e. $2\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq 4L_b(f(x) - \inf f)$ and $2\sigma_b^* = 2E[||\nabla f_B(x^*)||^2]$)

$\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$

2 Varianats on Stochastic Gradient Descent

2.1 Momentum

The key idea behind the momentum method is to incorporate a fraction of the previous update vector into the current update. Let's denote the previous update step as u_{t-1}. Then we can write the current update step u_t according to 37.

$$u^t = \alpha u^{t-1} + \gamma \nabla f(x^{(t)}) \tag{37}$$

where the $alpha \in (0, 1)$ is the momentum coefficient, which determines how much of the previous velocity is retained. This coefficient helps in accumulating a direction of persistent descent, smoothing over the updates. With this step, we can write the current iterate as,

$$x^{t+1} = x^t - u^t \tag{38}$$

.

The term u_t is also called the velocity term. The momentum term αu^{t-1} in the velocity term serves as a memory of past gradients:

- If gradients continue pointing in the same direction, the velocity grows in magnitude, allowing for faster convergence.
- If gradients change direction, the velocity's magnitude decreases, which helps mitigate oscillations and overshooting in steep regions of the parameter space.

This approach effectively dampens the oscillations and accelerates convergence towards the minimum of the loss function, particularly in landscapes where the surface curves more steeply in one dimension than in another. To visualize the effect of momentum in optimization let's imagine a ball rolling down a slope. If the slope does not have any turns the ball will keep accumulating velocity till it reaches the bottom. However, if there are turns the ball will slow down to navigate more efficiently.

NEED TO ADD A FIGURE ON MOMENTUM

2.2 Nestrov Accelerated Gradients (NAG)

NAG is also a momentum-based variant of SGD. The main difference between the momentum method and NAG lies in the gradient calculation stage. We have seen that in the momentum method, the update happens at x^t depending on the previous velocity v^{t-1} and the gradient of the function at x^t (37). In NAG the calculation of the gradient is done at a point ahead given by $\nabla f(x^t - \alpha u^{t-1})$.The intuition behind NAS is looking ahead and anticipating, which leads to better solutions. The update rule of NAS can be summarized as mentioned below.

1. Looking Ahead.

$$x^t_{lookahead} = x^t - \alpha v^{t-1} \tag{39}$$

2. Computing the gradient.

$$\nabla f(x^t - \alpha v^{t-1}) \tag{40}$$

3. Taking the gradient step.

$$\begin{aligned} x^{t+1} &= x^t_{lookahead} - \gamma \nabla f(x^t - \alpha v^{t-1}) \\ x^{t+1} &= x^t - \alpha v^{t-1} - \gamma \nabla f(x^t - \alpha v^{t-1}) \end{aligned} \tag{41}$$

This anticipatory step allows NAG to correct its course more responsively than standard Momentum, leading to potentially faster convergence and better handling of the curvature near optimal points. Essentially, NAG adds a level of foresight to updates, which can result in more efficient navigation of complex optimization landscapes.

NEED TO ADD FIGURE ON NESTROV

2.3 Ada Grad

AdaGrad is an adaptive learning rate method that modifies the general approach of gradient descent by allowing each parameter to have its own learning rate. This method addresses a common challenge in training machine learning models, where choosing an appropriate learning rate can be crucial for effective learning. Traditional gradient descent methods use a single learning rate for all parameters, which might not be optimal.

AdaGrad adjusts the learning rate for each parameter based on the history of gradients that have been computed for that parameter. This means that parameters associated with frequently occurring features will have their learning rates decreased, while parameters associated with infrequent features will have their learning rates increased. Such adjustments are beneficial because they make the model less sensitive to the scale of features and more responsive to each feature's specific behavior and importance. This feature-dependent scaling of the learning rate helps in dealing with data sparsity and enhances the convergence properties of the gradient descent optimization, particularly in complex models dealing with high-dimensional data.

Let's define, $f(x)$ to be the stochastic objective function with parameter x, the function evaluation at step t as $f_t(x)$, the gradient of the function with respect to x at step t to be $g_t(s)$. Further take,

$$\mathbf{G}_s = \sum_{t=1}^{s-1} g_t g_t^T \tag{42}$$

Now the update rule for Adagrad can be written as follows.

$$x_{t+1} = x_t - \gamma \mathbf{G}_t^{-\frac{1}{2}} g_t \tag{43}$$

A simplified version of the update rule can be written by only considering the diagonal elements of $\mathbf{G}$.

$$x_{t+1} = x_t - \gamma diag(\mathbf{G}_t)^{-\frac{1}{2}} g_t \tag{44}$$

This simplified version of the update step is computationally efficient when we are dealing with high-dimensional data. Additionally, to avoid the problems arise due to the matrix being singular,in practice a small offset is added to the diagonal elements of the matrix $\mathbf{G}$.

$$x_{t+1} = x_t - \gamma diag(\epsilon \mathbf{I} + \mathbf{G}_t)^{-\frac{1}{2}} g_t \tag{45}$$

Finally, let's look at the expanded version of the update rule.

$$\begin{bmatrix} x_{t+1}^{(1)} \\ x_{t+1}^{(2)} \\ \vdots \\ x_{t+1}^{(m)} \end{bmatrix} = \begin{bmatrix} x_t^{(1)} \\ x_t^{(2)} \\ \vdots \\ x_t^{(m)} \end{bmatrix} - \begin{bmatrix} \frac{\eta}{\sqrt{\varepsilon + G_t^{(1,1)}}} \\ \frac{\eta}{\sqrt{\varepsilon + G_t^{(2,2)}}} \\ \vdots \\ \frac{\eta}{\sqrt{\varepsilon + G_t^{(m,m)}}} \end{bmatrix} \odot \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(m)} \end{bmatrix}$$

Where $\odot$ is the Hadamard product between two matrices having the same dimensions. This provides a clear idea of how the per-parameter learning rate works. Here, γ is the paramter which describes the global learning rate. It must also be noted that as G accumulates, the learning rate slows down for each parameter and eventually no progress can be made, which is a weakness of Adagrad.

2.4 RMS Prop

2.5 Adam

3 Convergence Behavior of SGD and Minibatch SGD

3.1 SGD Convergence for Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a sequence of step sizes that satisfy $0 < \gamma_t < \frac{1}{4L_{max}}$.

Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_f^* \frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$

Proof:

Let us have $x^* \in \arg\min f$. We have already showed that when f is a Sum of Convex functions, $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$.

We know that, in Stochastic Gradient Descent, our iterates progress as such: $x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{i_t}(x^{(t)})$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma_t \nabla f_{i_t}(x^{(t-1)}) - x^*||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma_t \nabla f_{i_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t \nabla f_{i_t}(x^{(t-1)})\rangle + ||\gamma_t \nabla f_{i_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t \nabla f_{i_t}(x^{(t-1)})\rangle + \gamma_t^2||\nabla f_{i_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f_{i_t}(x^{(t-1)})\rangle + \gamma_t^2||\nabla f_{i_t}(x^{(t-1)})||^2$

Now, let's take the Expectation conditioned on $x^{(t-1)}$

$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

Based on the definition of convexity, we know that $f(y) \geq f(x) + \nabla(f(x))^T(y-x)$

This would mean that $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$

$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$

$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*)$

We can substitute this into the earlier equation we derived and get:

$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

Earlier, we proved that, when we have a function that is a sum of $L-$ Smooth functions and that is a sum of convex functions, $\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$

We can substitute this into the equations we derived:

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2(4L_{max}(f(x) - \inf f) + 2\sigma_f^*)$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2 4L_{max}(f(x) - \inf f) + 2\gamma_t^2\sigma_f^*$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 + (2\gamma_t)(2\gamma_t L_{max} - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_f^*$

Since $\gamma_t < \frac{1}{4L_{max}}$, $2\gamma_t L_{max} - 1 < \frac{-1}{2}$. We also know that $(f(x^{(t-1)}) - f(x^*)) > 0$ Hence

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - \gamma_t(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_f^*$$

Once again, let's take expectation over both sides of this inequality

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_f^*$$

$$\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2\sigma_f^*$$

$$\gamma_t\mathbb{E}(f(x^{(t-1)}) - \inf f) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2\sigma_f^*$$

Let's build this up recursively:
$\gamma_1\mathbb{E}(f(x^{(0)}) - \inf f) \leq \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(1)} - x^*||^2 + 2\gamma_1^2\sigma_f^*$

$$\gamma_2\mathbb{E}(f(x^{(1)}) - \inf f) \leq \mathbb{E}||x^{(1)} - x^*||^2 - \mathbb{E}||x^{(2)} - x^*||^2 + 2\gamma_2^2\sigma_f^*$$

$$\gamma_3\mathbb{E}(f(x^{(2)}) - \inf f) \leq \mathbb{E}||x^{(2)} - x^*||^2 - \mathbb{E}||x^{(3)} - x^*||^2 + 2\gamma_3^2\sigma_f^*$$

$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*$
We know that $\mathbb{E}||x^{(T)} - x^*||^2 > 0$. Hence, we can work with this inequality as such:

$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^* \leq \mathbb{E}||x^{(0)} - x^*||^2 + + \sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*$

$$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*$$

$$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq ||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*$$

Let's divide both sides of this inequality by $\sum_{t=1}^{T-1}\gamma_t$

$$\mathbb{E}[\sum_{t=1}^{T-1}\frac{\gamma_t}{\sum_{t=1}^{T-1}\gamma_t}(f(x^{(t-1)}) - f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t}$$

$$\mathbb{E}[\sum_{t=1}^{T-1}\frac{1}{\sum_{t=1}^{T-1}\gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t}$$

We know that f is convex. Hence, we can apply the Generalized Jensen's Inequality.
Based on the Generalized Jensen's Inequality, we can see that:

$f(\bar{x}^T) \leq \frac{1}{\sum_{t=1}^{T-1}\gamma_t}\sum_{t=1}^{T-1} y_t f(x^{(t-1)})$
Hence, our inequality becomes

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \mathbb{E}[\sum_{t=1}^{T-1}\frac{1}{\sum_{t=1}^{T-1}\gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t}$$

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a constant step size $\gamma_t = \gamma \leq \frac{1}{4L_{max}}$.
Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t = \frac{1}{\gamma T}\gamma\sum_{t=0}^{T-1} x^t = \frac{1}{T}\sum_{t=0}^{T-1} x^t$

Then for every $T \geq 1$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\gamma\sigma_f^*$

Proof. The proof of this is very simple as we did the majority of heavylifting in the last proof. We know that $\sum_{t=0}^{T-1} \gamma_t = \gamma T$ and $\sum_{t=0}^{T-1} \gamma_t^2 = T\gamma^2$. Let us substitute this into the last theorem and we will proceed from there.

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=1}^{T-1} \gamma_t}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + \frac{2\sigma_f^* T\gamma^2}{\gamma T}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\sigma_f^*\gamma$

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a vanishing step size $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ where $\gamma_0 \leq \frac{1}{4L_{max}}$

Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t$

Then for every $T \geq 1$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_f^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$

Proof. We know that our stepsize is decreasing. Hence, we can clearly see that $\gamma_t \leq \gamma_0 \leq \frac{1}{4L_{max}}$ for $t \geq 0$. Hence, we can apply the earlier result that we derived:

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + 2\sigma_f^* \frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t}$

Based on the Sum-Integral Bounds, we can say that $\sum_{t=0}^{T-1} \gamma_t = \gamma_0 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \frac{4\gamma_0}{5}\sqrt{T}$

$\sum_{t=0}^{T-1} \gamma_t^2 = \gamma_0^2 \sum_{t=1}^{T} \frac{1}{t} \leq 2\gamma_0^2 \log(T+1)$

Substituting it into the expected value, we get:

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_f^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$

Using the Sum-Integral Bounds, We will take a brief aside to discuss some theory that will be helpful to prove the theorem.

Let's say we have a function $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ that is decreasing.

It is clear to see that $\int_1^{T+1} f(x)\,dx \leq \sum_{x=1}^{T} f(x) = f(1) + \sum_{x=2}^{T} f(x) \leq f(1) + \int_1^T f(x)\,dx$

$\int_1^{T+1} f(x)\,dx \leq \sum_{x=1}^{T} f(x) \leq f(1) + \int_1^T f(x)\,dx$

Let's now use the functions $f(x) = \frac{1}{\sqrt{x}}$ and $f(x) = \frac{1}{x}$

Let's start with the function $f(x) = \frac{1}{\sqrt{x}}$

$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx \leq \sum_{x=1}^{T} \frac{1}{\sqrt{x}} \leq 1 + \int_1^T \frac{1}{\sqrt{x}}\,dx$

Let's start by working towards the lower bound
$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^{T+1} = 2\sqrt{T+1} - 2 = 2(\sqrt{T+1} - 1)$

We know that $\inf_{T\geq 1} \frac{\sqrt{T+1}-1}{\sqrt{T}} = \sqrt{2} - 1 > \frac{2}{5}$

$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^{T+1} = 2\sqrt{T+1} - 2 = 2(\sqrt{T+1} - 1) \geq \frac{4}{5}\sqrt{T}$

Now let's look at the upper bound:
$1 + \int_1^T \frac{1}{\sqrt{x}}\,dx = 1 + 2\sqrt{T+1} - 2 = 2\sqrt{T+1} - 1$
Combining both the Lower and Upper Bounds gives us:

$\frac{4}{5}\sqrt{T} \leq \sum_{x=1}^{T} \frac{1}{\sqrt{x}} \leq 2\sqrt{T+1} - 1$

Now let's analyze the other function $f(x) = \frac{1}{x}$

$\int_1^{T+1} \frac{1}{x}\,dx \leq \sum_{x=1}^{T} \frac{1}{x} \leq 1 + \int_1^T \frac{1}{x}\,dx$

Let's start by working towards the lower bound
$\int_1^{T+1} \frac{1}{x}\,dx = [\log(t)]_1^{T+1} = \log(T+1)$

Now let's look at the upper bound:
$1 + \int_1^T \frac{1}{x}\,dx = 1 + [\log(t)]_1^T = 1 + \log(T) \leq 2\log(T+1)$
Note: We know that $\sup_{T\geq 1} \frac{1+\log(T)}{\log(T+1)} \approx \sqrt{2} < 2$

$\log(T+1) \leq \sum_{x=1}^{T} \frac{1}{x} \leq 2\log(T+1)$

3.2 SGD Convergence for Strongly Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. We will also assume that f is p strongly convex. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ and we will assume that we have a constant stepsize satisfying $0 < \gamma < \frac{1}{2L_{max}}$. Then, we can say that for each iteration(i.e. $t \geq 0$)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1-\gamma p)^t ||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^* \tag{46}$$

Proof:
We know that, in Stochastic Gradient Descent, our iterates progress as such: $x^{(t+1)} = x^{(t)} - \gamma \nabla f_{i_t}(x^{(t)})$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma \nabla f_{i_t}(x^{(t-1)}) - x^*||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma \nabla f_{i_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{i_t}(x^{(t-1)})\rangle + ||\gamma \nabla f_{i_t}(x^{(t-1)})||^2$
$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{i_t}(x^{(t-1)})\rangle + \gamma^2 ||\nabla f_{i_t}(x^{(t-1)})||^2$
Now, let's take the Expectation conditioned on $x^{(t-1)}$
$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2 \mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

Based on the definition of strong convexity, we know that $f(y) \geq f(x) + \nabla(f(x))^T(y-x) + \frac{p}{2}||y-x||_2^2$

This would mean that $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$

$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$

$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2$

We can substitute this into the earlier equation we derived and get:

$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2 \mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2 \mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

Let's try to simplify $||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

$||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) - p\gamma||x^{(t-1)} - x^*||_2^2 + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

$(1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2$

Earlier, we proved that, when we have a function that is a sum of $L-$ Smooth functions and that is a sum of convex functions, $\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$

We can continue on with our proof as such:

$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2(4L_{max}(f(x^{(t-1)}) - \inf f) + 2\sigma_f^*)$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^*$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^*$

Since $\gamma < \frac{1}{2L_{max}}$, $2\gamma L_{max} - 1 < 0$. We also know that $\mathbb{E}(f(x^{(t-1)}) - f(x^*)) > 0$ Hence

$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^* \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + 2\gamma^2\sigma_f^*$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + 2\gamma^2\sigma_f^*$

Let's build this inequality recursively and see how it unfolds:
$\mathbb{E}||x^{(1)} - x^*||^2 \leq (1 - p\gamma)||x^{(0)} - x^*||^2 + 2\gamma^2\sigma_f^*$

$\mathbb{E}||x^{(2)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(1)} - x^*||^2 + 2\gamma^2\sigma_f^*$

$\mathbb{E}||x^{(2)} - x^*||^2 \leq (1 - p\gamma)((1 - p\gamma)||x^{(0)} - x^*||^2 + 2\gamma^2\sigma_f^*) + 2\gamma^2\sigma_f^*$

We can see that $\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^*$
Let's look at the term $\sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^*$.
$\sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^* < \sum_{n=0}^{\infty}(1 - p\gamma)^n 2\gamma^2\sigma_f^* = \frac{1}{p\gamma}2\gamma^2\sigma_f^* = \frac{2\gamma\sigma_f^*}{p}$
Hence, we can see that
$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^* \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^*$
$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^*$

3.3 Minibatch SGD Convergence for Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the Minibatch SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a sequence of step sizes that satisfy $0 < \gamma_t < \frac{1}{4L_b}$.
Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t$
$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_b^*\frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$
Proof:
Let us have $x^* \in \arg\min f$. We have already showed that when f is a Sum of Convex functions, $\sigma_b^* = \mathbb{V}[\nabla f_B(x^*)]$.
We know that, in Minibatch Stochastic Gradient Descent, our iterates progress as such: $x^{(t+1)} = x^{(t)} - \gamma_t\nabla f_{B_t}(x^{(t)})$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma_t\nabla f_{B_t}(x^{(t-1)}) - x^*||^2$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma_t \nabla f_{B_t}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t \nabla f_{B_t}(x^{(t-1)})\rangle + ||\gamma_t \nabla f_{B_t}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t \nabla f_{B_t}(x^{(t-1)})\rangle + \gamma_t^2 ||\nabla f_{B_t}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f_{B_t}(x^{(t-1)})\rangle + \gamma_t^2 ||\nabla f_{B_t}(x^{(t-1)})||^2$$

Now, let's take the Expectation conditioned on $x^{(t-1)}$

$$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$$

Based on the definition of convexity, we know that $f(y) \geq f(x) + \nabla(f(x))^T(y - x)$

This would mean that $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$$

$$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*)$$

We can substitute this into the earlier equation we derived and get:

$$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$$

Earlier, we proved that, when we have a function that is a sum of $L-$ Smooth functions and that is a sum of convex functions, $\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$

We can substitute this into the equations we derived:

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2(4L_b(f(x) - \inf f) + 2\sigma_b^*)$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2 4L_b(f(x) - \inf f) + 2\gamma_t^2\sigma_b^*$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 + (2\gamma_t)(2\gamma_t L_b - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_b^*$$

Since $\gamma_t < \frac{1}{4L_b}$, $2\gamma_t L_b - 1 < \frac{-1}{2}$. We also know that $(f(x^{(t-1)}) - f(x^*)) > 0$ Hence

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - \gamma_t(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_b^*$$

Once again, let's take expectation over both sides of this inequality

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_b^*$$

$$\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2\sigma_b^*$$

$$\gamma_t\mathbb{E}(f(x^{(t-1)}) - \inf f) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2\sigma_b^*$$

Let's build this up recursively:
$\gamma_1\mathbb{E}(f(x^{(0)}) - \inf f) \leq \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(1)} - x^*||^2 + 2\gamma_1^2\sigma_b^*$

$\gamma_2 \mathbb{E}(f(x^{(1)}) - \inf f) \leq \mathbb{E}||x^{(1)} - x^*||^2 - \mathbb{E}||x^{(2)} - x^*||^2 + 2\gamma_2^2 \sigma_b^*$

$\gamma_3 \mathbb{E}(f(x^{(2)}) - \inf f) \leq \mathbb{E}||x^{(2)} - x^*||^2 - \mathbb{E}||x^{(3)} - x^*||^2 + 2\gamma_3^2 \sigma_b^*$

$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*$
We know that $\mathbb{E}||x^{(T)} - x^*||^2 > 0$. Hence, we can work with this inequality as such:

$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^* \leq \mathbb{E}||x^{(0)} - x^*||^2 + + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*$

$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*$

$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq ||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*$

Let's divide both sides of this inequality by $\sum_{t=1}^{T-1} \gamma_t$

$\mathbb{E}[\sum_{t=1}^{T-1} \frac{\gamma_t}{\sum_{t=1}^{T-1} \gamma_t}(f(x^{(t-1)}) - f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$
$\mathbb{E}[\sum_{t=1}^{T-1} \frac{1}{\sum_{t=1}^{T-1} \gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

We know that f is convex. Hence, we can apply the Generalized Jensen's Inequality.
Based on the Generalized Jensen's Inequality, we can see that:

$f(\bar{x}^T) \leq \frac{1}{\sum_{t=1}^{T-1} \gamma_t} \sum_{t=1}^{T-1} y_t f(x^{(t-1)})$
Hence, our inequality becomes

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \mathbb{E}[\sum_{t=1}^{T-1} \frac{1}{\sum_{t=1}^{T-1} \gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$
$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t \in \mathbb{N}}$ with a constant step size $\gamma_t = \gamma \leq \frac{1}{4L_b}$.
Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t = \frac{1}{\gamma T} \gamma \sum_{t=0}^{T-1} x^t = \frac{1}{T} \sum_{t=0}^{T-1} x^t$

Then for every $T \geq 1$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\gamma \sigma_b^*$

Proof. The proof of this is very simple as we did the majority of heavylifting in the last proof. We know that $\sum_{t=0}^{T-1} \gamma_t = \gamma T$ and $\sum_{t=0}^{T-1} \gamma_t^2 = T\gamma^2$. Let us substitute this into the last theorem and we will proceed from there.
$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + \frac{2\sigma_b^* T\gamma^2}{\gamma T}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\sigma_b^* \gamma$

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t \in \mathbb{N}}$ with a vanishing step size $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ where $\gamma_0 \leq \frac{1}{4L_b}$
Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t$

Then for every $T \geq 1$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_b^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$

Proof. We know that our stepsize is decreasing. Hence, we can clearly see that $\gamma_t \leq \gamma_0 \leq \frac{1}{4L_b}$ for $t \geq 0$. Hence, we can apply the earlier result that we derived:

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + 2\sigma_b^* \frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t}$

Based on the Sum-Integral Bounds, we can say that $\sum_{t=0}^{T-1} \gamma_t = \gamma_0 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \frac{4\gamma_0}{5}\sqrt{T}$

$\sum_{t=0}^{T-1} \gamma_t^2 = \gamma_0^2 \sum_{t=1}^{T} \frac{1}{t} \leq 2\gamma_0^2 \log(T+1)$

Substituting it into the expected value, we get:

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_b^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$

3.4 Minibatch SGD Convergence for Strongly Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. We will also assume that f is p strongly convex. Let us say that the sequence of iterates generated by the Minibatch SGD Algorithm is $(x^{(t)})_{t \in \mathbb{N}}$ and we will assume that we have a constant stepsize satisfying $0 < \gamma < \frac{1}{2L_b}$. Then, we can say that for each iteration(i.e. $t \geq 0$)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t ||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_b^* \tag{47}$$

Proof:
We know that, in Minibatch Stochastic Gradient Descent, our iterates progress as such: $x^{(t+1)} = x^{(t)} - \gamma \nabla f_{B_t}(x^{(t)})$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma \nabla f_{B_t}(x^{(t-1)}) - x^*||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma \nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{B_t}(x^{(t-1)})\rangle + ||\gamma \nabla f_{B_t}(x^{(t-1)})||^2$
$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{B_t}(x^{(t-1)})\rangle + \gamma^2 ||\nabla f_{B_t}(x^{(t-1)})||^2$
Now, let's take the Expectation conditioned on $x^{(t-1)}$
$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Based on the definition of strong convexity, we know that $f(y) \geq f(x) + \nabla(f(x))^T(y - x) + \frac{p}{2}||y - x||_2^2$

This would mean that $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$

$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$

$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2$

We can substitute this into the earlier equation we derived and get:

$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Let's try to simplify $||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) - p\gamma||x^{(t-1)} - x^*||_2^2 + \gamma^2 \mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$(1-p\gamma)||x^{(t-1)}-x^*||^2 - 2\gamma(f(x^{(t-1)})-f(x^*)) + \gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)||x^{(t-1)}-x^*||^2 - 2\gamma(f(x^{(t-1)})-f(x^*)) + \gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Earlier, we proved that, when we have a function that is a sum of $L-$ Smooth functions and that is a sum of convex functions, $\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x)-\inf f) + 2\sigma_b^*$

We can continue on with our proof as such:

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2 - 2\gamma(f(x^{(t-1)})-f(x^*)) + \gamma^2(4L_b(f(x^{(t-1)})-\inf f) + 2\sigma_b^*)$

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2 + (2\gamma)(2\gamma L_b - 1)(f(x^{(t-1)})-f(x^*)) + 2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2 + (2\gamma)(2\gamma L_b - 1)\mathbb{E}(f(x^{(t-1)})-f(x^*)) + 2\gamma^2\sigma_b^*$

Since $\gamma < \frac{1}{2L_b}$, $2\gamma L_b - 1 < 0$. We also know that $\mathbb{E}(f(x^{(t-1)})-f(x^*)) > 0$ Hence

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2 + (2\gamma)(2\gamma L_b - 1)\mathbb{E}(f(x^{(t-1)})-f(x^*)) + 2\gamma^2\sigma_b^* \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2 + 2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2 + 2\gamma^2\sigma_b^*$

Let's build this inequality recursively and see how it unfolds:
$\mathbb{E}||x^{(1)}-x^*||^2 \leq (1-p\gamma)||x^{(0)}-x^*||^2 + 2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(2)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(1)}-x^*||^2 + 2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(2)}-x^*||^2 \leq (1-p\gamma)((1-p\gamma)||x^{(0)}-x^*||^2 + 2\gamma^2\sigma_b^*) + 2\gamma^2\sigma_b^*$

We can see that $\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^*$
Let's look at the term $\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^*$.
$\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^* < \sum_{n=0}^{\infty}(1-p\gamma)^n 2\gamma^2\sigma_b^* = \frac{1}{p\gamma}2\gamma^2\sigma_b^* = \frac{2\gamma\sigma_b^*}{p}$
Hence, we can see that
$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^* \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \frac{2\gamma}{p}\sigma_b^*$
$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \frac{2\gamma}{p}\sigma_b^*$

4 Algorithmic Stability of SGD

5 Numerical SGD Experiments

5.1 SGD: Ridge Regression

For the first set of experiments, Stochastic Gradient Descent will be used to solve the Ridge Regression Problem.